

비즈니스 규칙이 살아있는 정형 합성데이터 생성 솔루션

솔루션 소개서

- 2026년 공공데이터 제공 및 데이터기반행정 평가 지원
  - 가명정보 제공 및 합성데이터 개방 실적 가점(5)
  - 가명정보 제공 및 합성데이터 공유 실적 가점(5)
- 개방 및 공유를 위한 합성데이터 발굴 및 구축(생성) 지원

RealDataEcho

Executive Summary

RealDataEcho

규제로 막힌 데이터 활용

Problem

개인정보보호법 및 각종 규제로 인해 보유한 원본 데이터를 **외부 반출하거나, AI 학습에 직접 활용하는 것이 불가능**한 상황입니다.

합성데이터(Synthetic Data)

Solution

원본 데이터의 통계적 특성은 유지하면서, 개인정보는 포함하지 않는 가상 데이터를 생성하여 **자유로운 공유, AI 학습, 테스트**를 가능하게 합니다.

무결성 & 자동 검증

Key Value

단순 통계 유사도를 넘어 **온톨로지 기반 비즈니스 규칙 무결성**을 보장하며, 품질/프라이버시 자동 검증 리포트를 제공합니다.

폐쇄망 On-Premise

Deployment

금융/공공/의료 등 보안이 중요한 고객사를 위해 **데이터 외부 반출 없는 내부망 설치(On-Premise) 환경**을 기본 지원합니다.

고객의 현실: 데이터는 있는데, 쓸 수 없습니다

Ground Truth: From the Field

RealDataEcho

규제 강화

개인정보보호법

신용정보법

의료정보보호법

AI 데이터 수요 폭발

AI/ML 모델 훈련용

고품질 데이터수요 급증

ChatGPT · GenAI 확산

→ 모든 산업에서 AI 도입 가속

데이터 활용 제약에 따른 시장손실

익명화 시
데이터 유용성 감소율

62 %

가명정보 결합 평균 소요 기간

6개월

기업들이 개인정보 규제로
AI 프로젝트 연기/취소 경험

78%

심화되는 데이터 딜레마

규제 준수와 AI 혁신,
불가능해 보이는 양립이
필요한 시점

**새로운 데이터 접근
전략 필요**

비식별화만으로는 '활용'이 안 됩니다

기존 해법의 한계 vs 합성데이터의 가치

RealDataEcho



가명처리(블라인드)

활용성 제한

- ✓ 데이터 유용성 일부 유지
- ⚠ **재식별 리스크 잔존**
  (추가 정보 결합 시 위험)
- ✕ 외부 반출 및 공유 엄격 제한

ex. 주민번호 칸을 '000000-*******'와 같이 일부만 표시

VS



익명처리(삭제)

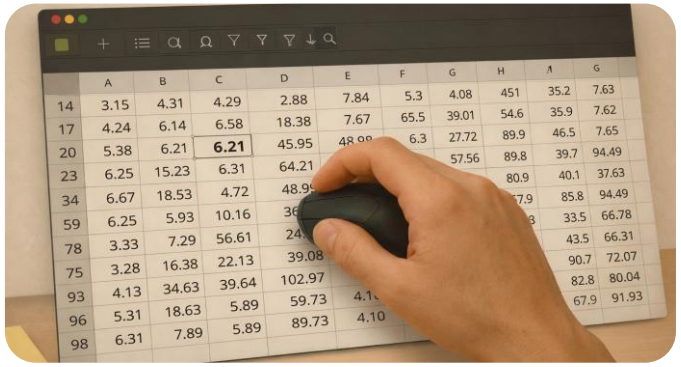
분석불가

- ✓ 재식별 리스크 최소화
- ↘ **유용성(Utility) 급감**
  (데이터 왜곡 심화)
- ✕ AI 학습 데이터로 부적합

ex. 주민번호 칸을 아예 없애는 형식

VS

BEST SOLUTION



합성데이터

완벽한 대안

- ✓✓ **원본 없이 통계적 특성 재현**
- 🛡 개인정보 유출 원천 차단
  (1:1 매칭 불가)
- **자유로운 공유·학습·테스트**

ex. 주민번호 등 실제 식별정보를 포함하지 않으면서, 원본 데이터와 유사한 통계적 특성을 가진 가상 데이터로 대체

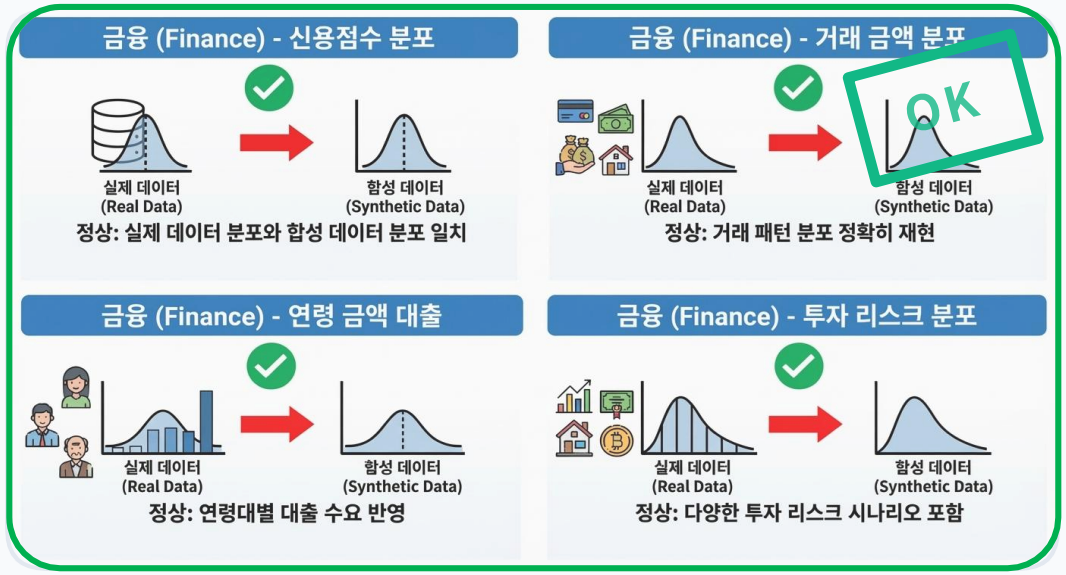
✓ 우리는 **'보안'**과 **'유용성'**을 동시에 만족시키는 합성을 제공합니다

숫자는 비슷해도, 현실 규칙이 깨지면 쓸 수 없습니다

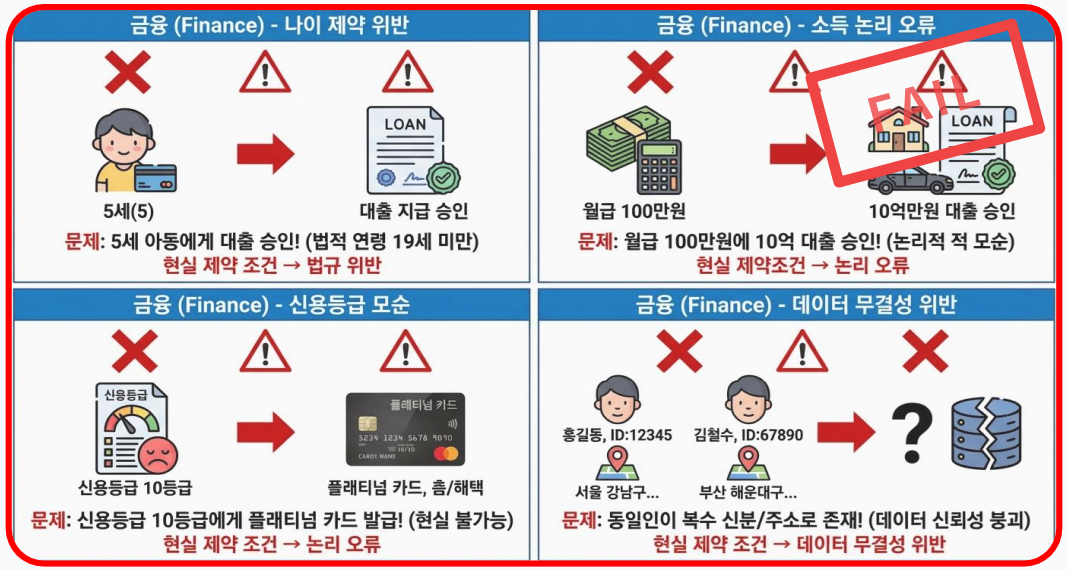
Problem: Statistical Similarity vs Business Rule Integrity

RealDataEcho

통계적 유사도 - 데이터 분포, 평균 등 통계적 특성은 원본과 유사



비즈니스 규칙 무결성 - 현실 규정, 법규 등을 위반한 데이터 생성



기존 합성데이터의 치명적 오류 사례

금융: DSR 규정 위반 "연소득 3천만원인데 대출원리금이 5천만원인 데이터 생성"
→ 현실 불가능, 모델 학습 시 노이즈 발생

재무: 계정관계 불일치 "자산 총계가 부채+자본 합계와 다른 재무제표 생성"
→ 회계 기본 원칙 위배, 분석 신뢰도 0점

> 규칙이 깨진 데이터는 AI와 분석가의 신뢰를 즉시 잃게 만듭니다.

정우마루 합성데이터 솔루션

Product Overview: 3-Layer Integrated Architecture

RealDataEcho

정우마루 합성데이터 솔루션

논리적 무결성 검증 + 합성데이터 생성 솔루션

통계를 넘어 논리로

- 현실 구조(Reality Structure) 설계
- 원본 데이터 없이 AI 학습 데이터 생성
- 논리적 무결성 보장

도메인 지식 (온톨로지)

통계 패턴

20대 + 강덕 40년 오류 차단

고정밀 합성데이터

경량 LLM-RAG 기반 자가학습

지능형 자동화
- sLLM + RAG → 규정·매뉴얼 분석
- 도메인 규칙 자동 추출

확장성 확보
- SI 인력 의존 탈피
- 구축 기간·비용 대폭 절감

산업별 적용 및 혁신 가치

의료	금융	제조
(환자 데이터, 신약 개발)	(사기 탐지, 리스크 관리)	(품질 예측, 공정 최적화)

기대 효과

데이터 확보 제약 해소, AI 모델 신뢰성 향상, 시장 출시 기간 단축, 법적 리스크 최소화

논리 검증 및 규제 준수 통합

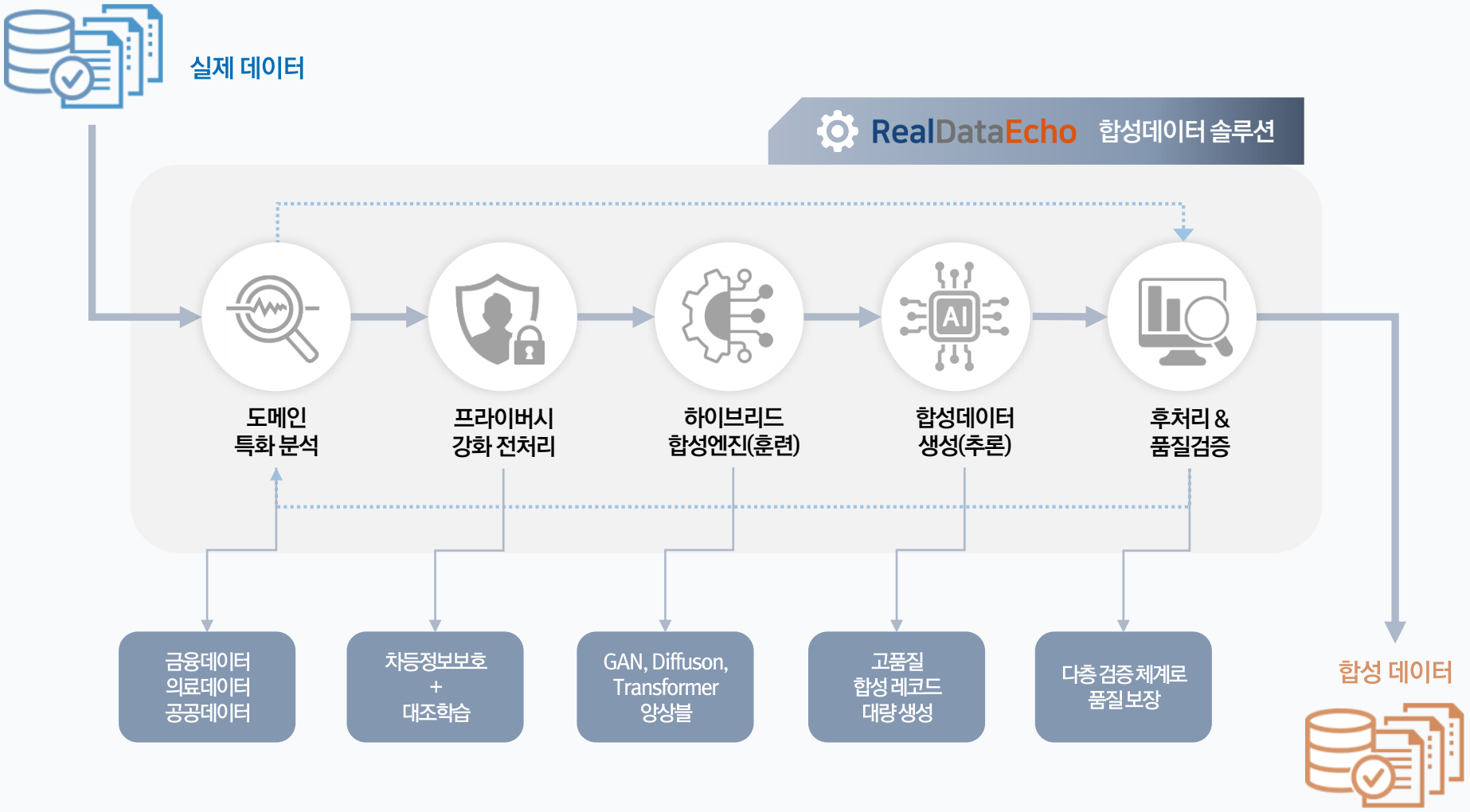
규제 정보 수집 (Global/Local) → 온톨로지 기반 규칙 엔진 → 합성데이터 생성 (개인정보 비식별화) → 논리적/윤리적 무결성 검증 → 최종 안전 데이터셋 제공

☑ 글로벌 표준 (GDPR, CCPA 등) 준수 및 데이터 무결성 보증

정형데이터 특화 5단계 통합 파이프라인

End-to-End Workflow for Structured Data Synthesis

RealDataEcho



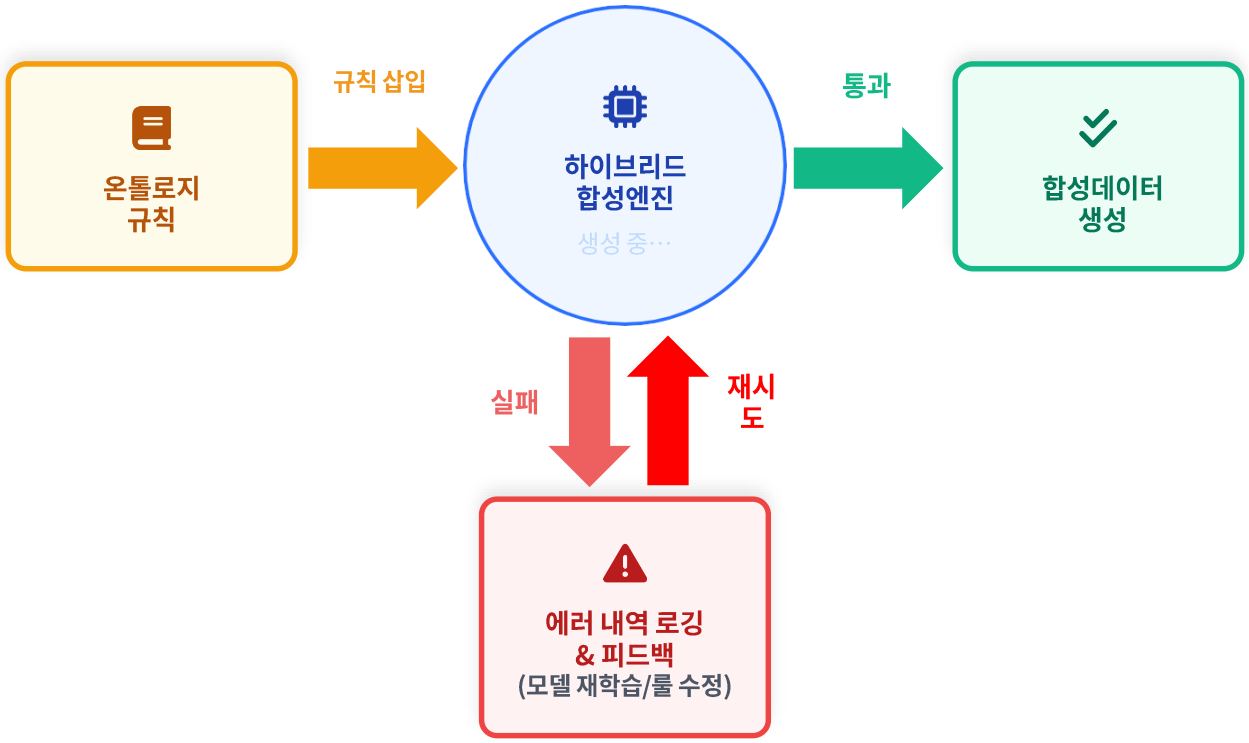
온톨로지 규칙으로 '생성 시점'에 오류를 차단

Integrity Enforcement with Hard Constraints

RealDataEcho

Rule-Based Generation Loop

생성 데이터의 규칙 위배 시, 해당 규칙과 오류 데이터 출력 및 피드백 후 재시도



Hard Constraint란?

단순 통계적 확률에 의존하지 않고, **도메인 규칙을 강제 조건으로 설정**합니다. 규칙 위반 데이터는 생성 단계에서 즉시 폐기하거나 재생성합니다.

적용 규칙 예시

회계/재무

계정관계규칙 (Accounting Rule)

자산 = 부채 + 자본, 차변/대변 합계 일치

금융 규제

DSR 및 LTV 한도 규정

소득 대비 대출 원리금 비율 상한 준수

자격 요건

기간 및 가입 자격 논리

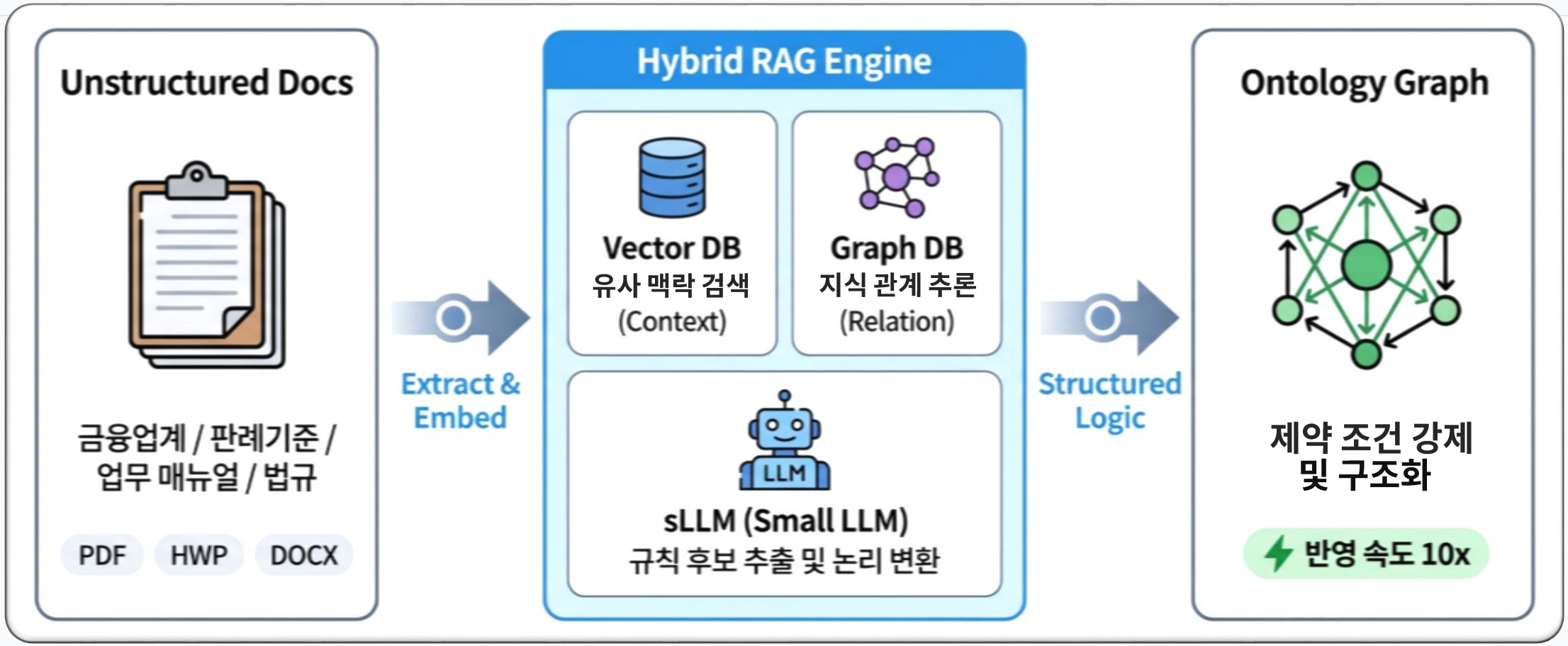
입사일 < 퇴사일, 미성년자 상품 가입 불가 등

위반 레코드 0건 보장

규정/매뉴얼에서 규칙을 빠르게 추출·갱신

Accelerated Rule Extraction with sLLM & Hybrid RAG Architecture

RealDataEcho



폐쇄망 온프레미스 구동 (On-Premise Ready)

외부 API 호출 없이, 내부 서버에 경량 모델(sLLM)을 직접 설치하여 데이터 유출 원천 차단

데이터 외부 반출 없이, 내부에서 독립 구동

Secure On-Premise Deployment within Closed Network

RealDataEcho

고객사 내부 네트워크 (폐쇄망)

폐쇄망 환경 지원 시스템



원본데이터 입력

원본 데이터 업로드 및 데이터 스키마 자동 검증



내부 서버 접속

폐쇄망 온프레미스 환경 내 데이터 처리



도메인 규칙 추출 (온톨로지)

제약조건, 비즈니스 룰 자동 추출



합성데이터 및 검증 리포트 생성

원본과 통계적, 논리적 동등한 합성데이터 생성 및 검증

하드웨어 권장 사양

CPU	RAM	GPU	Storage
10 Core+	32GB+	NVIDIA 4090	1TB NVMe

※상기 권장 사양은 고객사 환경에 맞추어 변경 가능합니다.

온프레미스 설치 기본 제공

고객사 서버실 또는 프라이빗 클라우드에 직접 설치하여 데이터의 외부 유출 가능성이 적습니다.

외부 연결 차단 (폐쇄망)

인터넷 연결이 없는 폐쇄망 환경에서도 라이선스 인증 및 모델 학습/생성이 가능합니다.

엔터프라이즈 보안 기능

접근 통제(RBAC), 감사 로그(Audit Log), 데이터 암호화 등 기업 보안 정책을 준수합니다.

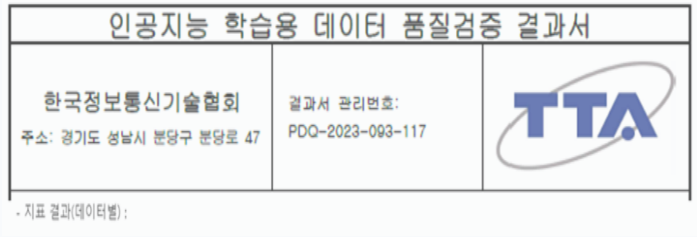
데이터 유출 리스크 Zero화

Evaluation Protocol (측정 기준)

Standardized Methodology for Synthetic Data Verification

RealDataEcho

정확도 (Data Accuracy)



인공지능 학습용 데이터 품질검증 결과서

한국정보통신기술협회
주소: 경기도 성남시 분당구 분당로 47

결과서 관리번호:
PDQ-2023-093-117

TTA

- 지표 결과(데이터별):

구 분	JSD(0.4 이하)	pMSE(0.2 이하)	Corr.Diff(0.7 이하)
개인 신용평가정보	0.003	0.061	0.036
기업 신용평가정보	0.003	0.004	0.033

Jensen-Shannon Divergence (JSD) : 두 확률 분포 간의 유사성을 측정, 0에 가까우면 두 분포가 유사함을 의미
Pairwise Mean Squared Error (pMSE) : 두 데이터 세트 간의 차이를 측정, 값이 낮을수록 모델의 성능이 좋다는 것을 의미
Correlation Difference : 두 데이터 세트 간의 변수 쌍(pair of variables)의 상관 관계의 변화 또는 차이를 설명

통계적 유사성

JSD 0.003

최저 수준 달성



성능지표 (Downstream performance metrics)

AI 모델	원본 테스트 결과	합성 테스트 결과	테스트 결과 유사도
개인 부도율 예측	0.599	0.582	98%
기업 부도율 예측	0.781	0.753	97%

AI 학습 효율성

97~98%

원본 대비 학습효율



개인정보보호 (Data Privacy)

데이터	Record linkage	TCAP
개인 CB 데이터	▪ 원본과 일치하는 레코드 없음	TCAP 평균 0.12로 안전

프라이버시 안전성

TCAP 0.12

재식별 위험 없음



신용보증기금: 기업재무제표 합성 PoC (2개월)

Financial Statement Synthesis Proof of Concept

RealDataEcho

PoC 완료

KODIT 신용보증기금

본 프로젝트 수행 예정

데이터 범위

기업 속성 데이터 (Firm Attributes)

업종코드 | 종업원수 | 설립연도 | 보증유무 | 외감여부

재무제표 데이터 (Financial Statements)

- 재무상태표 (BS) & 손익계산서 (IS)
  - 장 (Chapter)
  - 관 (Section)
  - 항 (Item)
  - 목 / 세목 (Detail)

핵심 성과

100%

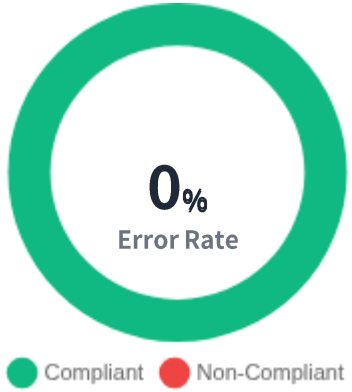
Accounting Rule Consistency

차대변 일치 / 자산=부채+자본 등
회계 등식 무결성 확보

98%

Sector Pattern Retention

업종별 재무 비율 패턴 유지
(제조업 vs 서비스업 특성 반영)



왜 이 결과가 중요한가요?

- 기존 모델은 통계적 분포만 학습하여 **회계 등식이 깨진 데이터**를 생성했습니다.
- 정우마루는 **온톨로지 제약조건(Hard Constraint)**을 적용하여, 수백만 건 생성 시에도 단 한 건의 위반 없이 완벽한 정합성을 달성했습니다.

> **"재무 데이터는 규칙이 한 번이라도 깨지면 0점입니다.
> 정우마루는 그 0점의 가능성을 원천 차단했습니다."**

적용 시나리오

Diverse Use Cases: From AI Training to External Collaboration

RealDataEcho

01

AI 모델 학습/고도화

부족한 학습 데이터를 증강하거나 불균형을 해소합니다.
(신용리스크, 부도예측, 이상탐지 모델 등)

모델 성능 15~20% 향상

02

시스템 테스트/부하테스트

실제 고객 데이터 패턴을 반영한 대용량 테스트셋 생성.
(차세대 시스템 오픈, 거래 패턴 재현 등)

개인정보 유출 위험 Zero

03

외부 협업 (Open Innovation)

민감 정보 없이 안전하게 데이터를 외부로 공유합니다.
(스타트업, 컨설팅, 연구기관, 해커톤 등)

협업 착수 시간 6개월 → 1주 단축

04

공공 데이터 개방/민간 혁신

활용도가 높은 고품질 마이크로 데이터를 대국민 개방.
(금융 마이데이터, 공공 빅데이터 포털 등)

데이터 가치 극대화 및 민간 활용

실제 구동 화면과 리포트

Operational Pipeline, Verification Dashboard, and Integrity Logs

RealDataEcho

JeongWooMaru Studio > New Project

01. Generation Pipeline

Upload — Pre-process — 3 Config & Train — 4 Generate

Table Schema

Financial_Statements_2025

Privacy Budget (Epsilon)

Medium (ε = 1.0)

Ontology Rule Set

Finance_Standard_v2.0

Generation Mode

Hard Constraint (Strict)

Training Generator Model... 75%

Est. Time: 2m 30s GPU Usage: 82%

Report Viewer

02. Validation Dashboard

98.5%	Low	100%
Utility Score	Privacy Risk	Integrity





Rule Execution Log

03. Integrity Log

Rule ID	Constraint Type	Status	Violations
R-ACC-01	Asset = Liab + Equity	PASS	0
R-FIN-05	Revenue >= Net Income	PASS	0
R-BIZ-12	Emp. Count > 0	PASS	0
R-DSR-03	DSR Limit < 40%	PASS	0

All 24 Hard Constraints Satisfied.

고객 데이터는 외부 반출 없이 내부 인프라(On-Premise)에서 안전하게 처리됩니다.

도입 프로세스: PoC → 구축 → 운영

Implementation Roadmap & Key Deliverables

RealDataEcho

01 PoC (검증)

6 ~ 8 Weeks

- 샘플 데이터 및 요구사항 분석
- 핵심 온톨로지 규칙(Rule) 정의
- 합성데이터 생성 및 검증 리포트

02 규칙/온톨로지 규칙 설정

6 ~ 8 Weeks

- 도메인 전체 테이블로 범위 확장
- 복합 규칙 및 예외 케이스 정의
- 커스텀 생성 모델 파라미터 튜닝

03 온프레미스 설치/연동

2 ~ 4 Weeks

- 사내 폐쇄망 서버 솔루션 설치
- 레거시 DB 및 파이프라인 연동
- 보안성 검토 및 권한 설정

04 운영 및 고도화

상시 (Ongoing)

- 정기적 모델 재학습(Update)
- 신규 규칙 반영 및 최신화
- 품질/무결성 리포트 자동화

PoC Deliverables

PoC 종료 시 산출물

합성 데이터셋
Synthetic Dataset (CSV/DB)

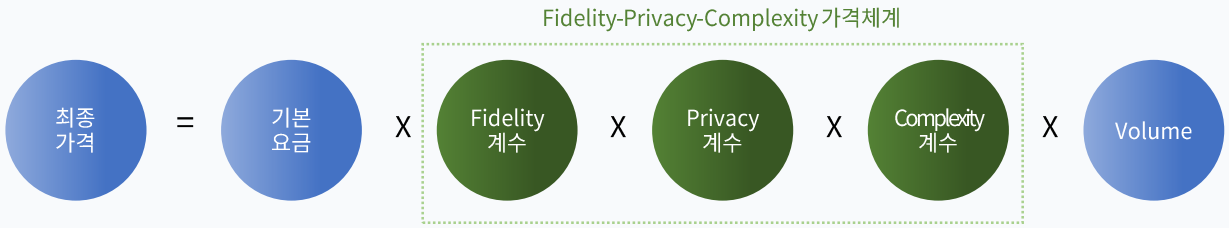
품질 검증 리포트
Quality & Privacy Report

합리적 계층형 과금 (Fidelity-Privacy-Complexity)

Flexible Pricing Model Based on Data Sensitivity & Scale

RealDataEcho

정우마루 과금모델 가격체계



구분	Fidelity (통계적 유사도)	Privacy (보안 강도)	Complexity (데이터복잡도)
주요내용	• 원본 데이터와 얼마나 같은가? • 90% vs 95%는 분석 결과가 완전히 달라짐	• 법적 리스크 직결 • 금융/의료 등 민감 데이터는 최고 수준 보안 필요	• 합성데이터는 컬럼수 증가할수록 생성난이도 기하급수적으로 증가
측정방법	JSD, Corr Differnce, pMSE	• k-익명성 (k=5, 10, 50 선택) • 차등 프라이버시 (ε값 조정)	• 컬럼수, 제약조건
예시	• 리스크 모델링 → 95% 이상 필수 • 단순 테스트 → 85%면 충분	• 외부 공유용 → 최고 수준 • 내부 분석용 → 중간 수준	• simple : 컬럼수 20개 이하 • Medium : 21-50 컬럼 • complex : 51-100 컬럼 • Enterprize : 100~200 컬럼 • Mega : 200 이상

정우마루 과금모델 가격체계 예시

	POC	Enterprise On-Premise	Cloud SaaS
타겟고객	모든 고객	금융기관 / 공공기관 / 의료기관(제약회사)	중소/중견, 스타트업
가격(Price)	3천만원	1억원(base)	월간 1백만원 ~
Fidelity (통계정확도)	85%	95%	90%
Privacy (보안강도)	Medium ★★★☆☆ 신원노출위험 < 0.3	Extreme ★★★★★ 신원노출위험 < 0.01	High ★★★★☆ 신원노출위험 < 0.1
Complexity (복잡도)	컬럼갯수 30개 이하	컬럼갯수 200개 이하	컬럼갯수 100~120개
생성규모	10만건 이하	무제한	100만건 / 월
계약기간	3개월 이하	1년 + 연간유지보수(15%)	월간 또는 연간 구독

정우마루가 선택받는 이유

Why Leading Enterprises Choose JeongWooMaru

RealDataEcho

비교 항목 (Key Features)	정우마루	기타 솔루션 (Cloud-first / General)
정형데이터 최적화 관계형 DB 구조 및 시계열 특성 반영	✓	이미지/텍스트 중심 (정형 성능 낮음)
온프레미스/폐쇄망 기본 제공 데이터 외부 반출 없는 내부 구축	✓	SaaS Cloud 중심 (반출 필수)
규칙 무결성 (Hard Constraint) 비즈니스 로직 위반 0% 보장	✓	통계적 유사도만 고려 (규칙 위반 발생)
다운스트림 성능 검증 AI 모델 학습 시 원본 대비 성능 유지(AUC 등)	✓	✓
프라이버시 정량 평가 Record Linkage, TCAP 등 재식별 위험 측정	✓	일부 지원
검증 리포트/감사 대응 데이터 품질 및 무결성 증빙 자동 생성	✓	미지원

> 단순 생성기가 아닌, **검증 가능한 데이터 생산 시스템**입니다.

공공/금융에서 검증 중인 기술

Proven Technology in Public & Financial Sectors

RealDataEcho

2026

- 신용보증기금 기업 재무제표 합성데이터 생성 PoC 수행

2025

- TIPS 프로그램 선정 (R&D 과제 수행)
- 창업 중심대학 과제 선정 및 수행
- 벤처기업 혁신성장유형 인증

2024

- AI 학습용 금융 합성데이터 생성 시스템 외 1건 특허 등록
- 기업부설 연구소 인증
- 정우마루 법인 설립 (24.09)

MAIN REF

KODIT 신용보증기금

기업 재무제표 합성데이터 생성 PoC

기간	**2개월**
데이터	**기업 재무제표 데이터**
핵심기술	**금융 합성데이터 생성, 온톨로지 제약 조건**

계정관계 정합성 100% 달성

특허 등록 2건
합성데이터 생성 및 검증 관련

벤처기업 인증
혁신성장유형

기업부설연구소
KOITA 인정

- **2026년 공공데이터 제공 및 데이터기반행정 평가 지원용 S/W**
  - 가명정보 제공 및 합성데이터 개방 실적 가점(5)
  - 가명정보 제공 및 합성데이터 공유 실적 가점(5)
- **개방 및 공유를 위한 가명·합성데이터 발굴 지원 컨설팅**

망설이지 마시고, PoC로 확인하시죠.

귀사 내부망(On-Premise)에서,
귀사 데이터로 직접 생성하고,
품질·프라이버시·무결성 리포트까지 제공합니다.

Enterprise Security Ready
데이터 외부 반출 Zero, 내부망 완벽 구동

01 **PoC 미팅 & 요구사항 정의**
합성 목적 및 데이터 스키마 확인

02 **샘플 데이터 / 룰 문서 수령**
보안 절차에 따른 데이터 이관 (내부망)

03 **PoC 수행 및 결과 리포트**
합성데이터 산출물과 검증 보고서 제출

RealDataEcho

제조사
주식회사 정우마루

공공부문 Distributor
주식회사 엠아이티소프트

Email
saels@mitsoft.kr

Phone
02-719-0380